PO.DAAC Metadata Best Practices and Vision Beyond 2015

ESIP Data Stewardship Committee October Teleconference
19 October 2015

David Moroni
Jet Propulsion Laboratory
Pasadena, CA

These activities were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Outline

I. What is PO.DAAC?
II. Variety of Data Holdings
III. Variety of Data Formats
IV. Metadata Standards
V. Metadata Harvesting
    I. Dataset/Collection
    II. Granule
VI. Supporting Interoperability
VIII. Completeness and Conformity
IX. Beyond 2015
X. Conclusions
XI. Questions/Feedback

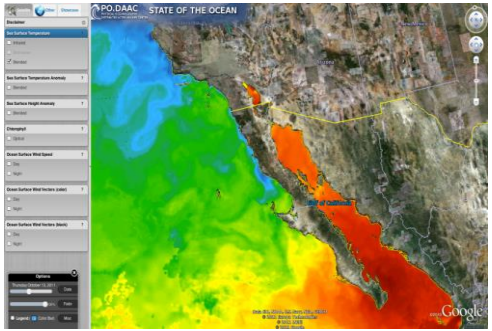
What is PO.DAAC?

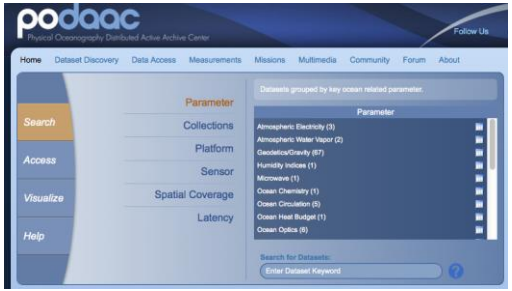
Data Management & Stewardship

Preserve NASA's data for the benefit of future generations

Data Access

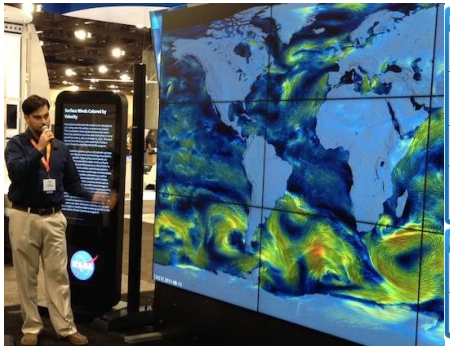
Provide intuitive services to discover, select, extract and utilize data





Science Information Services

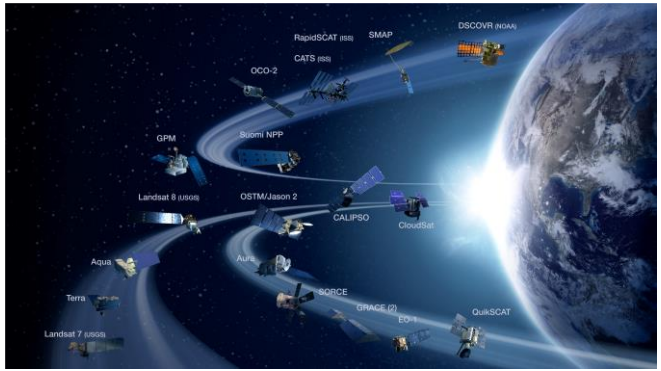
Provide a knowledgebase to help a broad user community understand and interpret satellite ocean data and related information



Variety of Data Holdings

NASA Missions & Projects

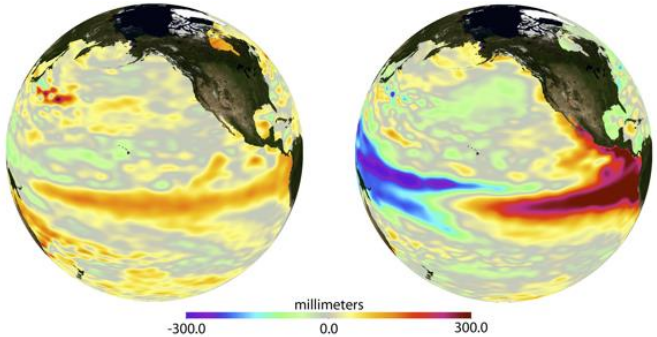
Seasat, TOPEX/Poseidon, Jason-1, NSCAT, SeaWinds on ADEOS-II, QuikSCAT, GRACE, GHRSST, MEaSUREs, Aquarius, SPURS, ISS-RapidScat, CYGNSS (2016), GRACE-FO (2017), SWOT (2020)



Ocean & Climate Community Driven

Value-added datasets in support of NASA programs

Gravity, Ocean Circulation & Currents, Ocean Surface Salinity, Ocean Surface Topography, Ocean Vector Winds, Sea Surface Temperature, *Ocean Color, Sea Ice*



Variety of Data Formats

- Interoperable and Self-Describing:
  - netCDF-3, netCDF-4 (classic and enhanced), HDF-4, HDF-5
  - 75% of our public-domain datasets
- Legacy (i.e., outdated; limited interoperability)
  - ASCII, flat binary, GIF, JPEG, PNG
  - 25% of our public-domain datasets
- Overall, both science data users and data producers are increasingly aware and becoming more preferential toward self-describing and interoperable data formats.

Metadata Standards

- What we currently adhere to:
  - Dataset/Collection: ISO 19115, ISO 8601, GCMD*, ECHO/REVERB*
  - Granule/File: CF, ACDD, ISO 8601, GHRSST GDS-2*
- What we are planning to implement:
  - Consolidated Metadata Repository* (CMR)
  - ISO 19157
- Who we are in communication with:
  - ESIP: CF Standards, Information Quality, Interoperability and Information Technology, Preservation and Stewardship
  - ESDSWG: Citations, Data Quality, Interoperability, Provenance, Search Relevancy

* Denotes NASA-proprietary standards

Metadata Harvesting: Dataset/Collection

- Submission Agreement (a.k.a MOU, Inter-Project Agreement)
  - document to kick off our ability to ingest new datasets and metadata
  - provides preliminary collection/dataset metadata
- Data Management and Archiving System (DMAS):
  - Compliant with ISO 19115, GMCD, and ECHO/REVERB
    - will eventually support CMR
  - Supports SQL-based backend query, modification, and extraction
- Data Management Tool (DMT):
  - Web-based GUI to submit dataset/collection metadata
  - Interfaces directly with DMAS

Metadata Harvesting: Granule

- Stored in DMAS
  - enables flexible collection and harvesting of metadata in real-time
  - required metadata collection artifacts are specified using DMT
- Data Handler
  - an automated tool that interfaces with a web-crawler to search, extract, and archive the data granules/files and corresponding metadata
  - both generic and custom-configurations are available
  - can be used in both “push” and “pull” data ingest modes

Supporting Interoperability

- Web Services: http://podaac.jpl.nasa.gov/ws/
- OPenDAP: http://opendap.jpl.nasa.gov/
- THREDDS: http://thredds.jpl.nasa.gov/
- Webification (w10n): http://podaac-w10n.jpl.nasa.gov/
- Advocating use of CF and ACDD, thus enabling extended interoperability:
  - Panoply, THREDDS, and many APIs that rely on THREDDS servers

Supporting Data Quality

- ISO 19157 is being scoped out by ESDIS and PO.DAAC.
  - Addresses data quality at the dataset/collection-level.
  - Functions as a container for data quality attributes and data quality variables.
  - Recommended by NASA Data Quality Working Group and industry experts.
- CF quality flag attributes address data quality at the granule level.
- GHRSST GDS-1 and GDS-2 both address data quality at the granule level.
  - Standardized quality flagging for all SST datasets.
  - Standardized uncertainty estimates at the pixel level.
- Data Granule/File integrity validation via MD5 Checksum
- Documentation: User Guides, READMEs, References to Refereed Literature
- User Forums: https://podaac.jpl.nasa.gov/forum/

Completeness and Conformity

- Metadata Compliance Checker (MCC):
  - http://podaac-uat.jpl.nasa.gov/mcc/
  - Automated granule/file metadata completeness and compliance checking
  - ACDD, CF, ISO-8601, GHRSST GDS-2
  - Supports manual single-file upload and OPeNDAP URLs
  - Output options: HTML or PDF
  - Still in Beta state with further development being scoped out
- Web Portal and DMT:
  - Functions as a visual validation tool to ensure dataset/collection metadata is complete and consistent for all PO.DAAC datasets.
  - DMT is used to make quick edits in real-time.
- GCMD, ECHO/REVERB:
  - Functions as an external validation tool to ensure both completeness and standards compliance to ISO 19115 as well as NASA-proprietary GCMD and ECHO/REVERB standards.

Beyond 2015

- Online dataset/collection metadata submission form for data producers.
  - Point of entry for new datasets to be archived and distributed.
  - Compatible with DMAS and DMT.
  - DMT would be used to by DAAC Data Engineers to make final edits.
- Granule/file metadata editor:
  - Leveraging NCO utilities package to edit data variable and global attributes.
  - Provides retroactive and non-intrusive metadata completeness and compliance to ACDD, CF, ISO-8601.
  - Leveraging MCC backend to validate metadata completeness and compliance.
  - Would help to bring “dark data” into the modern age of provenance and interoperability.

Conclusions

- More users are preferring interoperable, self-describing data formats.
- Data Producers have only recently started catching up with this user demand, hence why many require some tutelage with standards:
  - MCC provides a nice entry point at validating and understanding metadata completeness and compliance at granule level.
- With increased self-description comes the need for standardized metadata.
  - CF, ACDD, and ISO provide thorough guidance in this regard.
  - CF and ACDD go a step beyond ISO to ensure broad and long-term interoperability.
- DAACs and other data centers may need non-invasive metadata editing tools to transform “dark” datasets into the modern-era of interoperability.
  - Data Producers may also find such tools to be quite useful: e.g., NOAA CDR, MEaSUREs

Questions/Feedback

- Data Best Practices:
  - https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices